Project Title- IPL DATA ANALYSIS USING PIG AND HIVE

Scope:

The Indian Premier League (IPL), also known as [TATA](https://en.wikipedia.org/wiki/Tata_Group) IPL for sponsorship reasons, is a men's [T20](https://en.wikipedia.org/wiki/T20_cricket) franchise cricket league of [India](https://en.wikipedia.org/wiki/India). It is annually contested by ten teams based out of seven Indian cities and three Indian states. The league was founded by the [Board of Control for Cricket in India](https://en.wikipedia.org/wiki/Board_of_Control_for_Cricket_in_India) (BCCI) in 2007. [Brijesh Patel](https://en.wikipedia.org/wiki/Brijesh_Patel) is the incumbent chairman of IPL. It is usually held annually in summer across India between March to May and has an exclusive window in the [ICC Future Tours Programme](https://en.wikipedia.org/wiki/ICC_Future_Tours_Programme).

The IPL is the most-attended cricket league in the world and in 2014 was ranked sixth by average attendance among all sports leagues. In 2010, the IPL became the first sporting event in the world to be broadcast live on [YouTube](https://en.wikipedia.org/wiki/YouTube).

Over the course of its run starting from its inaugural season in 2008 till the recently concluded one in 2022, there have been various winners with the franchises of Chennai and Mumbai winning the title multiple times.

In this project we have taken the IPL data from the years 2008 to 2016 and have applied the concepts of Hadoop and implemented in hive and pig.

Sample Dataset:

Two cricket data files with Indian Premier League data from 2008 to 2016 is used as a data source. The files are as follows: 1.matches.csv – Provides details about each match played

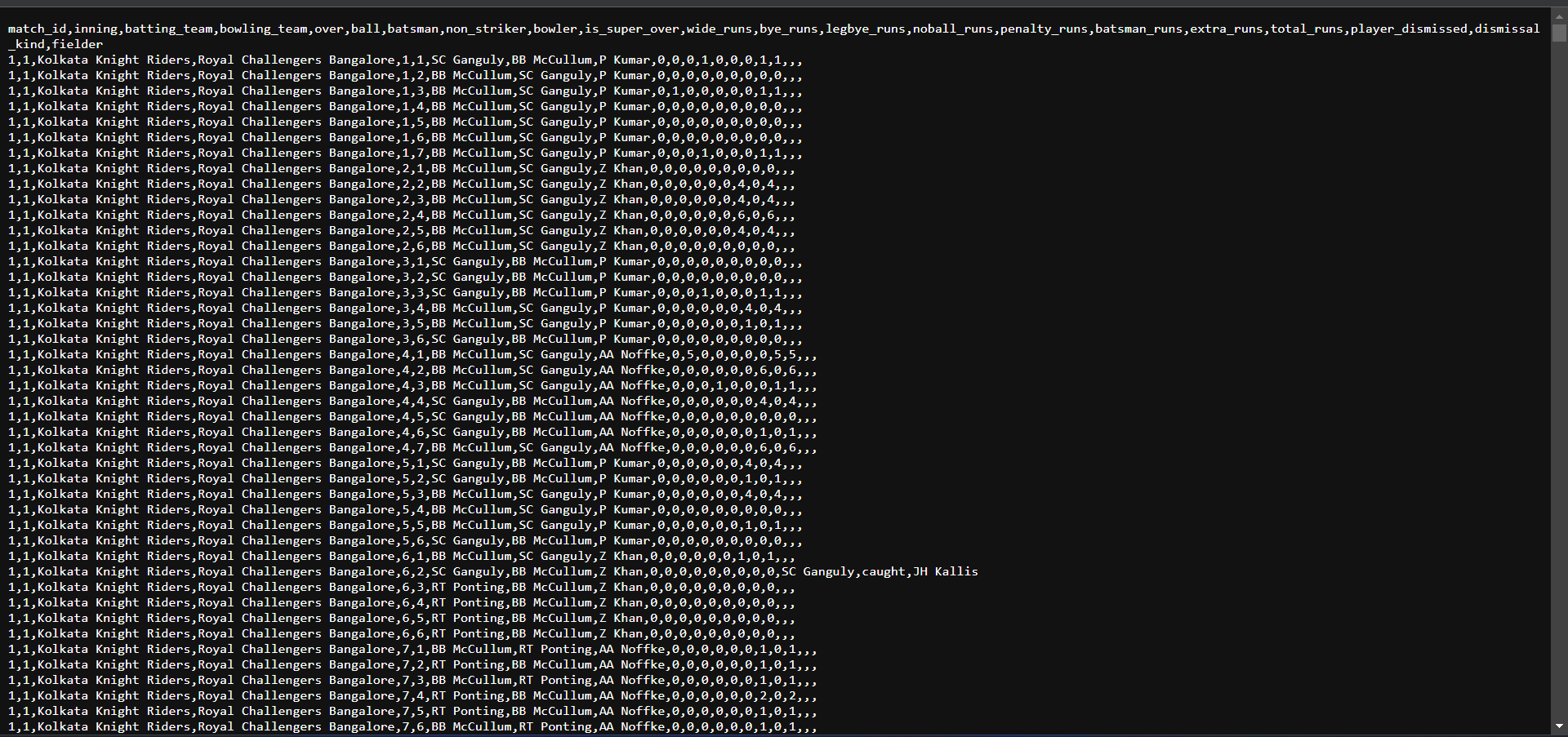
2.deliveries.csv – Provides details about consolidated deliveries of all the matches

These files are extracted and loaded into Hive. The data is further processed, transformed, and analyzed to get the winner for each season and the top 5 batsmen with maximum run in each season and overall season.

AIM:

To find the information of certain players and to find the winner of the IPL editions from 2008 to 2016 on the two given datasets and analyse them using pig and hive in Hadoop.

Deliveries File:



The deliveries file has the attributes of match\_id, innings, batting team, bowling\_team, over, ball, batsman, non striker, bowler, super\_over, wide\_runs, bye\_runs, legbye\_runs, no\_ball\_runs, penalty\_runs, batsman\_run, extra\_runs, total\_run, player\_dismissed, dismissal\_kind, fielder.

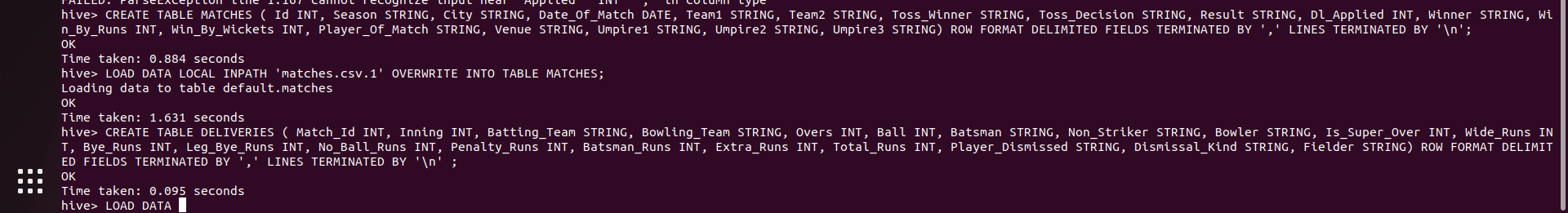
Matches File:



The matches dataset contains the attributes id, season, city, date, team1, team2, toss\_winner, toss\_decision, dl\_applied, winner, win\_by\_runs.

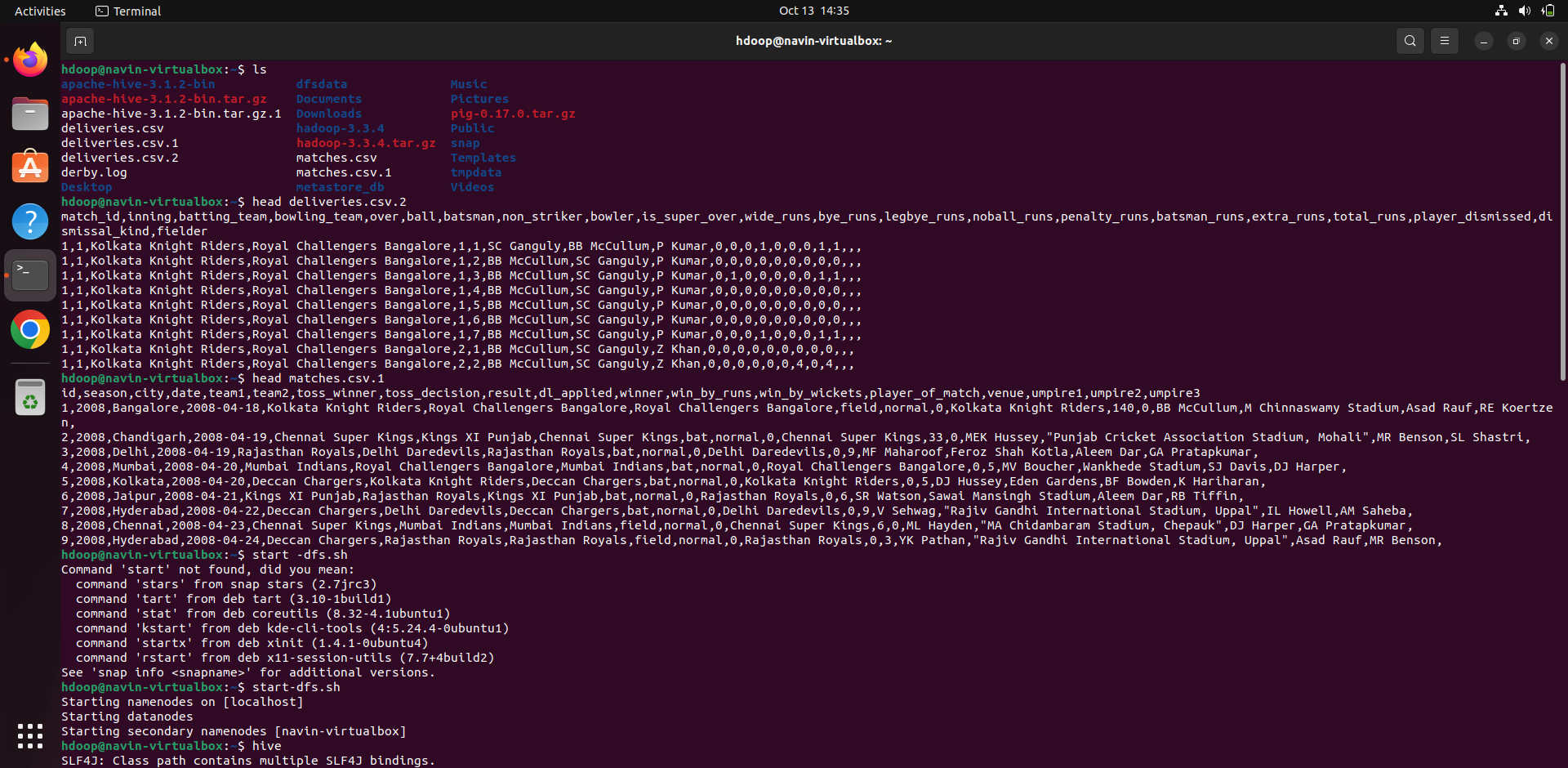
HIVE QUERIES:

Creating table for matches and deliveries



Displaying contents of the table

HEAD MATCHES.CSV



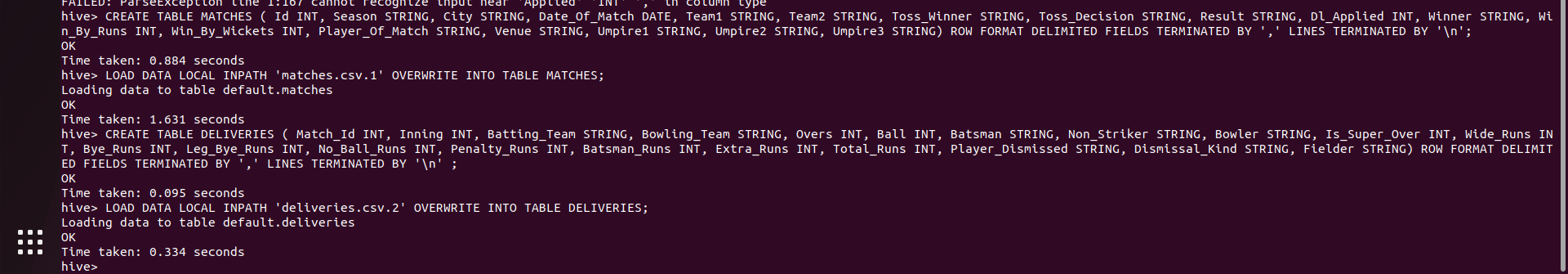
HEAD DELIVERIES.CSV



Storing the Data as a database and loading the data saved in the hdfs directory.

Code:

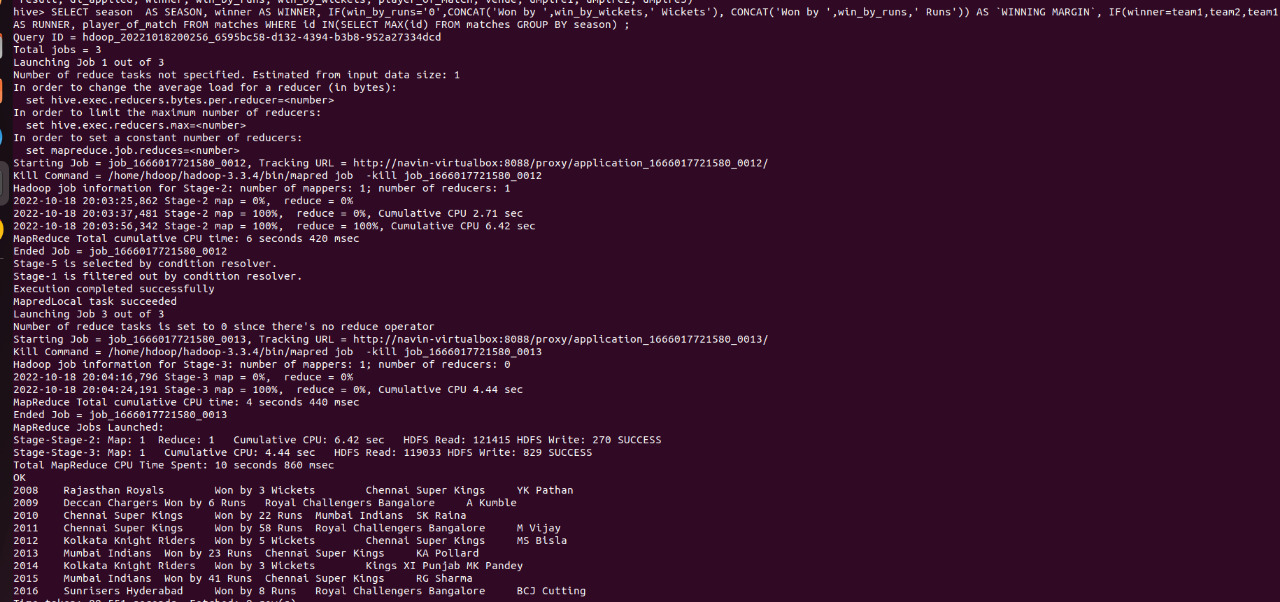
CREATE TABLE MATCHES



Query to select the winner of each season

CODE:

SELECT season AS SEASON, winner AS WINNER, IF(win\_by\_runs='0', CONCAT('Won by ',win\_by\_wickets, 'Wickets'), CONCAT('Won by ',win\_by\_runs,' Runs')) AS `WINNING MARGIN`, IF(winner=team1,team2,team1) AS RUNNER, player\_of\_match FROM matches WHERE id IN(SELECT MAX(id) FROM matches GROUP BY season) ;

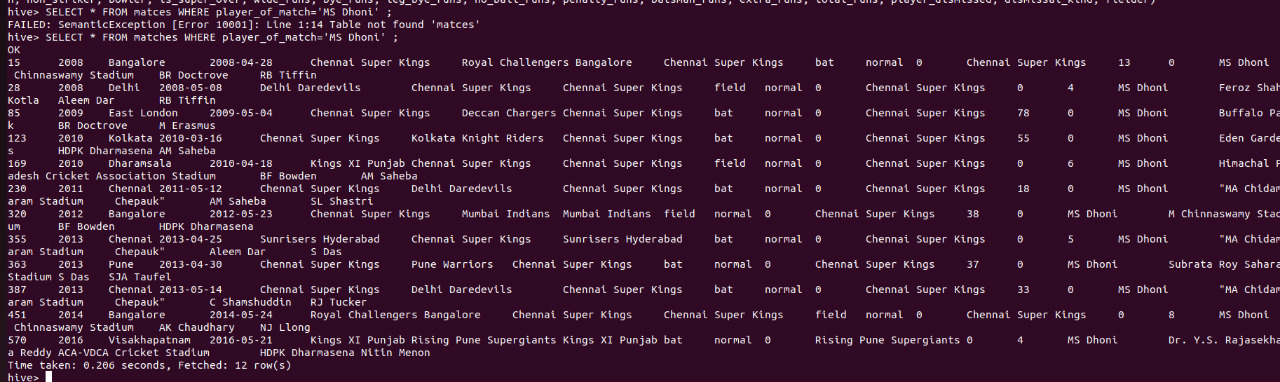


Query to get info on MS Dhoni

Code:

SELECT \* FROM MATCHES WHERE PLAYER\_OF\_MATCH=’MS DHONI’ ;

OUTPUT:

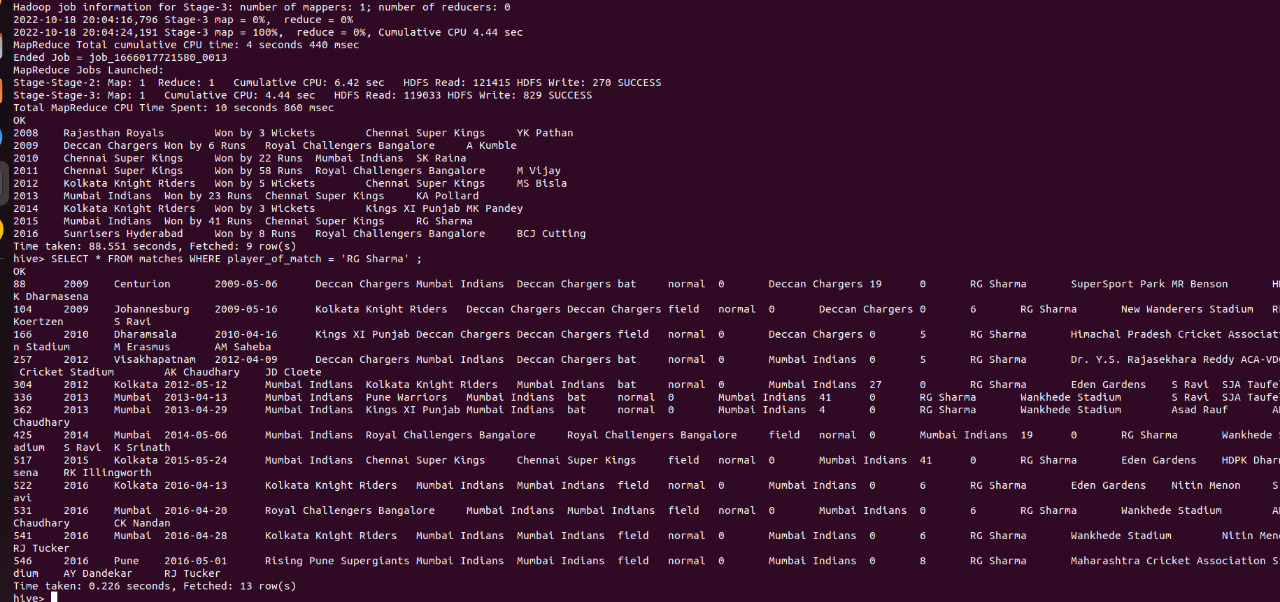


Query to get info on Rohit Sharma

CODE:

SELECT \* FROM MATCHES WHERE PLAYER\_OF\_MATCH=’RG SHARMA’;

OUTPUT:

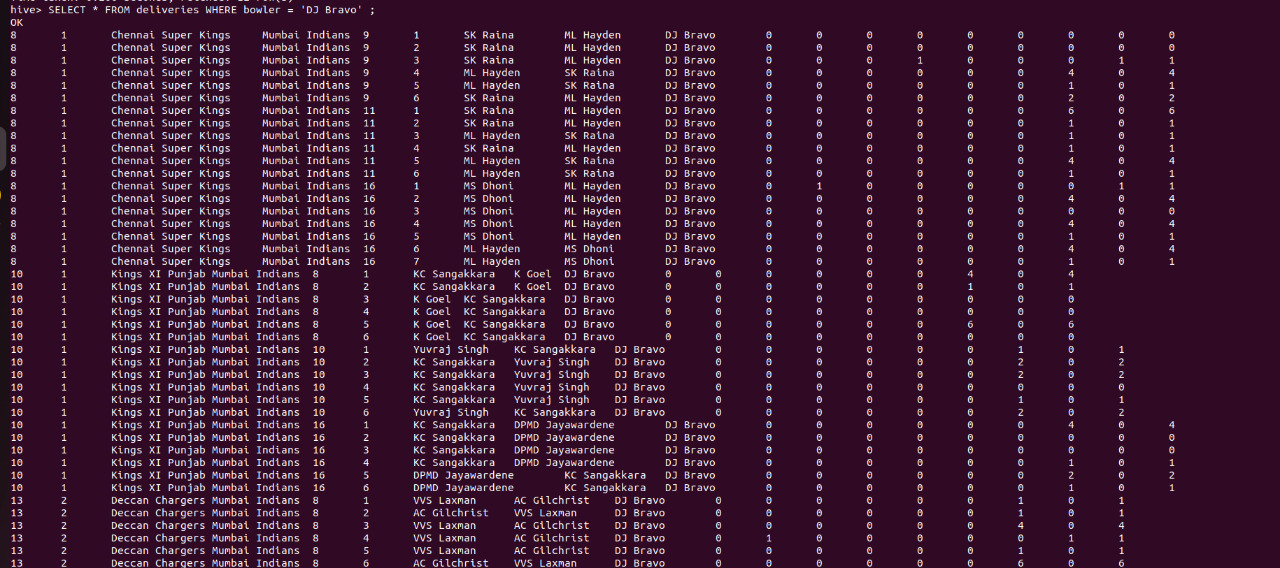


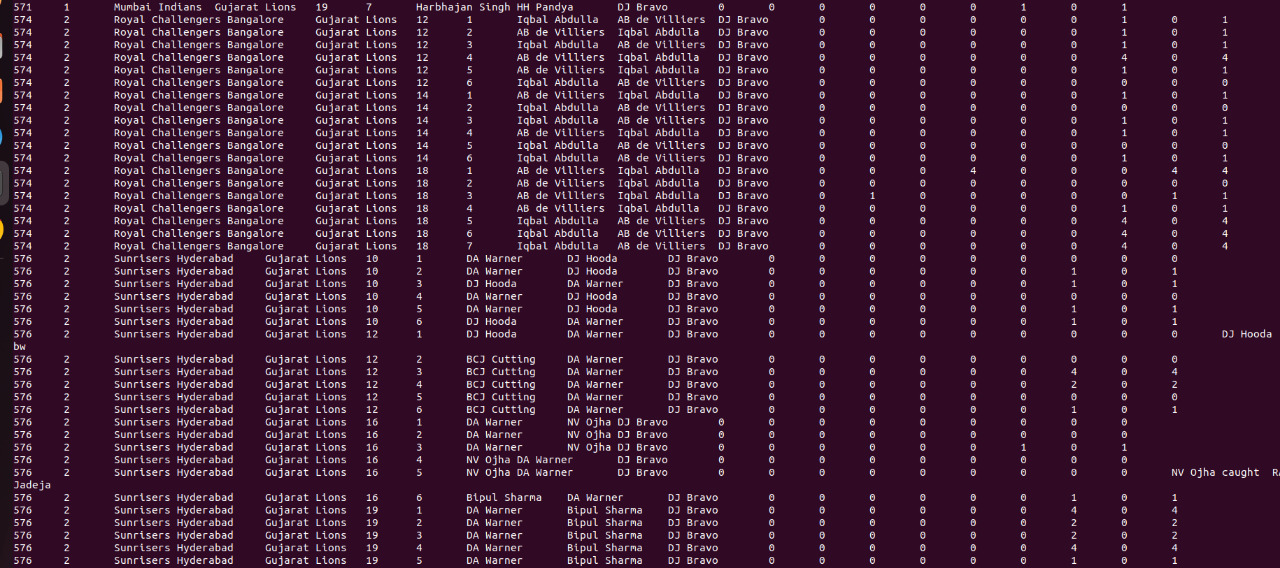
From Deliveries table

Query to fetch detail for bowler DJ Bravo

CODE:

SELECT \* FROM DELIVERIES WHERE BOWLER=’DJ BRAVO’;

OUTPUT: 

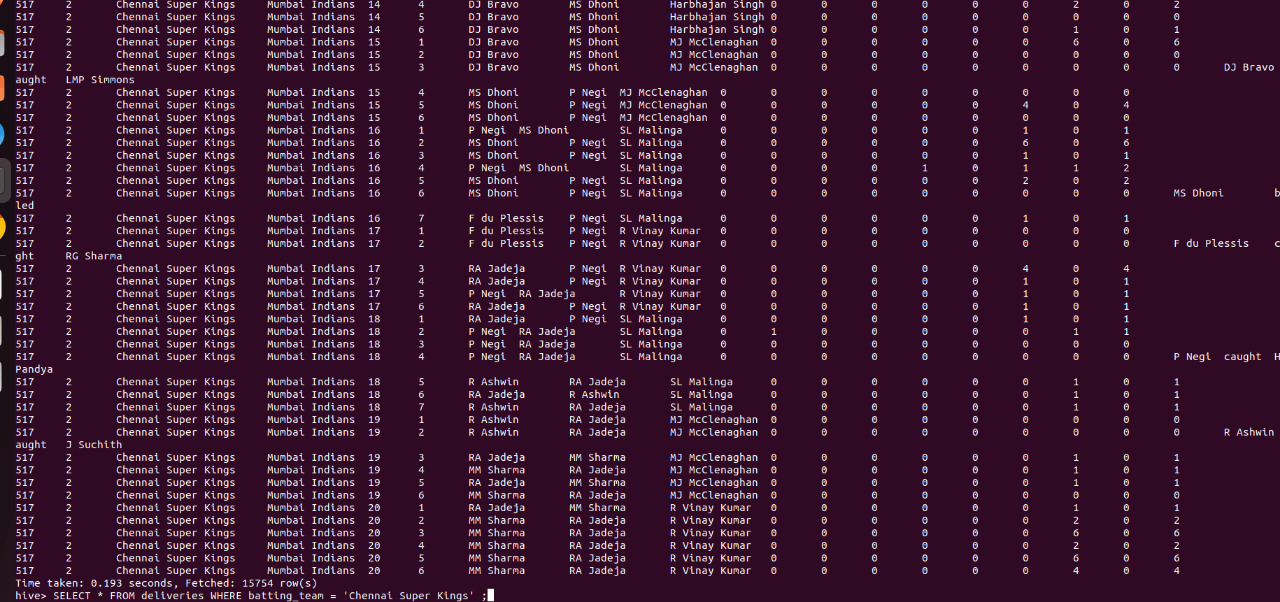


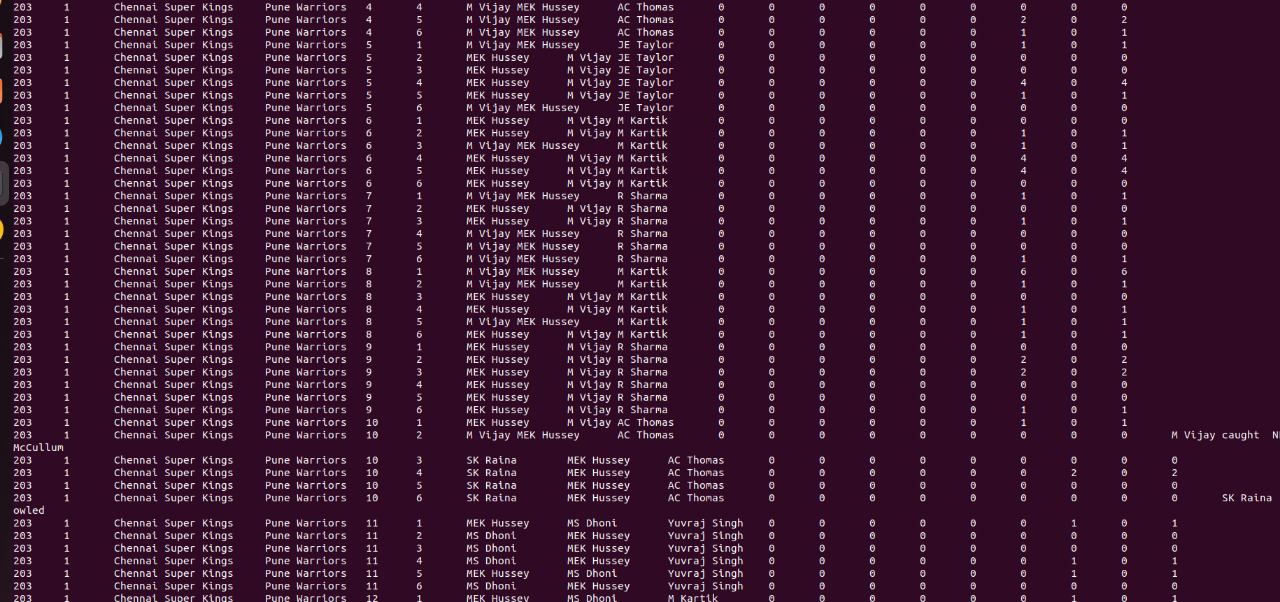
Query to get info on the team Chennai Super Kings

CODE:

SELECT \* FROM DELIVERIES WHERE BATTING\_TEAM=’CHENNAI SUPER KINGS’ ;

OUTPUT:



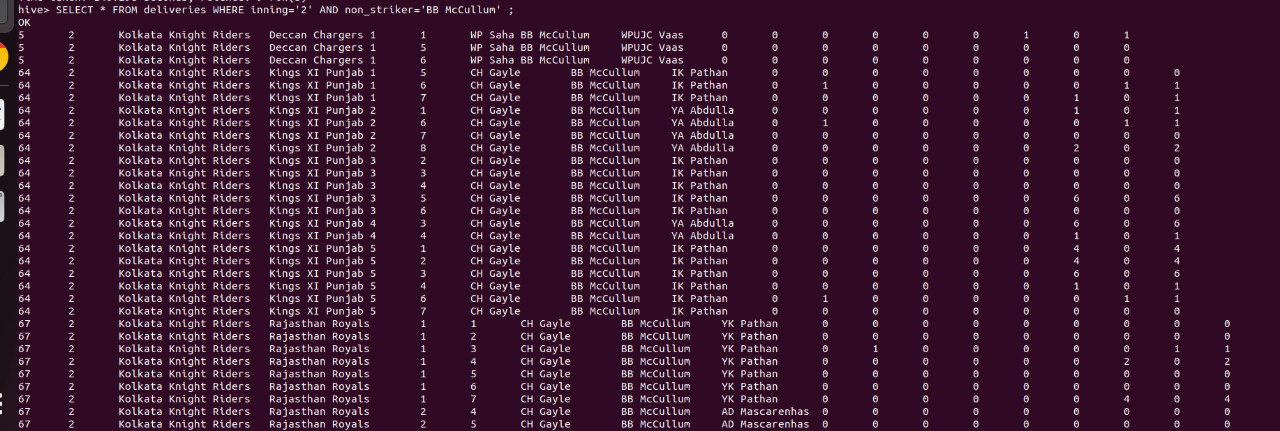


Query to get info on non striker batsman

CODE:

SELECT \* FROM deliveries WHERE inning=’2’ AND non\_striker=’BB McCullum’;

OUTPUT:

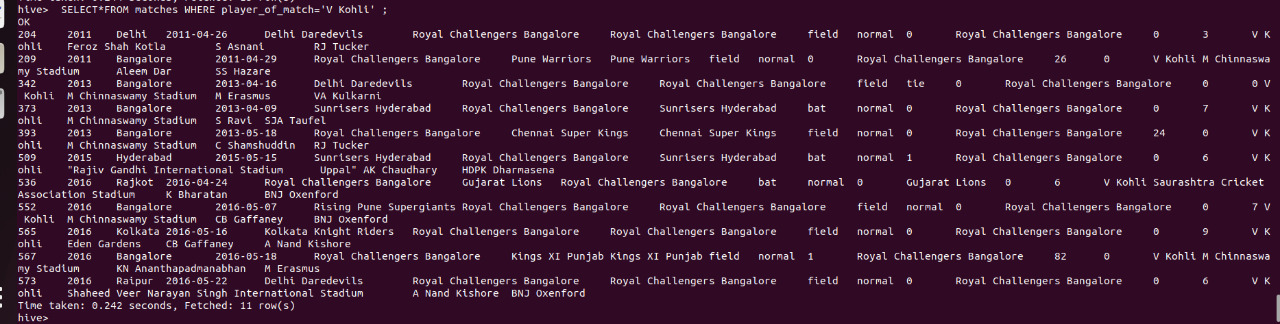


Query to get info on Virat Kohli

CODE:

SELECT \* FROM matches WHERE player\_of\_match=’V Kohli’;

OUTPUT:



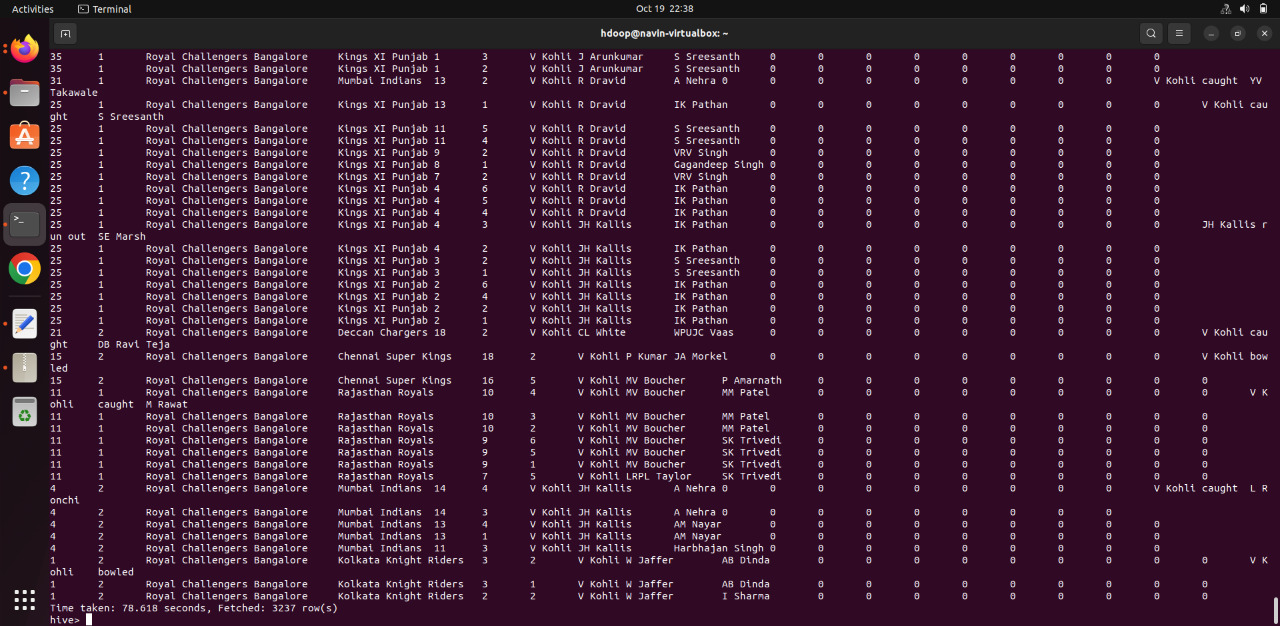
ORDER BY CLAUSE:

CODE:

SELECT \* FROM MATCHES WHERE batsman=’V Kohli’ ORDER BY total\_runs DESC;

OUTPUT:



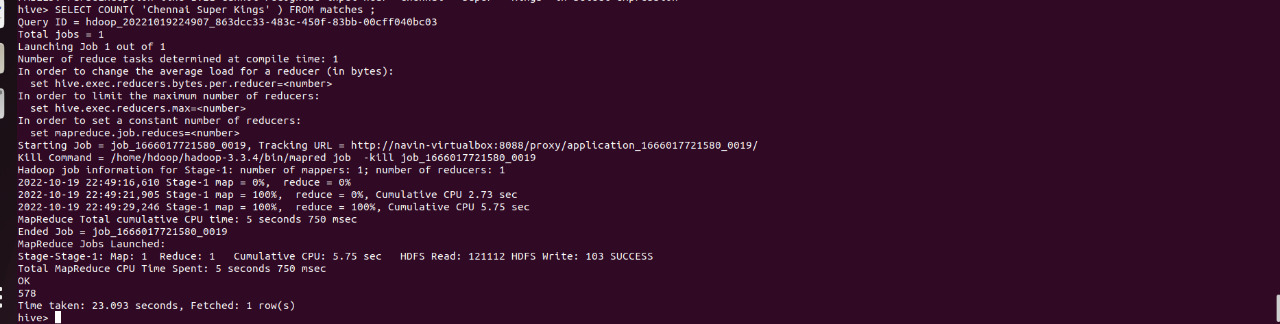


Query to get count of a team

CODE:

SELECT COUNT(‘Chennai Super Kings’) FROM matches;

OUTPUT:



Query to get the count of matches played between two teams throughout the seasons

CODE:

SELECT COUNT(‘Chennai Super Kings’) FROM matches WHERE team1=’Chennai Super Kings’ AND team2=’Mumbai Indians’ ;

SELECT COUNT(‘Royal Challengers Bangalore’) FROM matches WHERE team1=’Kolkata Knight Riders’ AND team2=’Royal Challengers Bangalore’ ;

OUTPUT:

